Relation Extraction

Heavily based on slides from:
Prof. Sameer Singh

CS 295: STATISTICAL NLP
WINTER 2017

February 23, 2017
Outline

- Introduction to Relation Extraction
- Hand-written Patterns
- Supervised Machine Learning
- Semi and Unsupervised Learning
Outline

- Introduction to Relation Extraction
- Hand-written Patterns
- Supervised Machine Learning
- Semi and Unsupervised Learning
Relation and Event Extraction

Perhaps the most important application of practical NLP can do many useful things.

Far from being a solved problem,

it touches almost all areas of language processing.
Goal: “machine reading”

- Acquire structured knowledge from unstructured text
Information extraction

• IE = extracting information from text
• Sometimes called *text analytics* commercially
• Extract *entities*
  o People, organizations, locations, times, dates, prices, ...
  o Or sometimes: genes, proteins, diseases, medicines, ...
• Extract the *relations* between entities
  o Located in, employed by, part of, married to, ...
• Figure out the *larger events* that are taking place
Information extraction

- IE = extracting information from text
- Sometimes called *text analytics* commercially

**Extract entities**
- People, organizations, locations, times, dates, prices, ...
- Or sometimes: genes, proteins, diseases, medicines, ...

**Extract the relations** between entities
- Located in, employed by, part of, married to, ...

**Figure out the larger events** that are taking place
Information extraction

• IE = extracting information from text
• Sometimes called *text analytics* commercially
• Extract entities
  o People, organizations, locations, times, dates, prices, ...
  o Or sometimes: genes, proteins, diseases, medicines, ...
• Extract the *relations* between entities
  o Located in, employed by, part of, married to, ...
• Figure out the *larger events* that are taking place

Relation and Event Extraction
Knowledge Extraction

Extract a comprehensive set of relations from text.

John was born in Liverpool, to Julia and Alfred Lennon.
Machine-readable summaries

textual abstract: summary for human

structured knowledge extraction: summary for machine
More applications of IE

- Building & extending knowledge bases and ontologies
- Scholarly literature databases: Google Scholar, CiteSeerX
- People directories: Rapleaf, Spoke, Naymz
- Shopping engines & product search
- Bioinformatics: clinical outcomes, gene interactions, …
- Patent analysis
- Stock analysis: deals, acquisitions, earnings, hirings & firings
- SEC filings
- Intelligence analysis for business & government
Relation Extraction

Company report: “International Business Machines Corporation (IBM or the company) was incorporated in the State of New York on June 16, 1911, as the Computing-Tabulating-Recording Co. (C-T-R)…”

Extracted Complex Relation:

- **Company-Founding**
  - Company: IBM
  - Location: New York
  - Date: June 16, 1911
  - Original-Name: Computing-Tabulating-Recording Co.

But we will focus on the simpler task of extracting relation **triples**

- Founding-year(IBM, 1911)
- Founding-location(IBM, New York)
Extracting Relation Triples

The Leland Stanford Junior University, commonly referred to as Stanford University or Stanford, is an American private research university located in Stanford, California... near Palo Alto, California... founded the university in 1891.
Where does the set of relations come from?

[discuss]
ACE corpus and competitions

Automated Content Extraction

- PERSON-SOCIAL
  - Family
  - Business
  - Lasting Personal
- PHYSICAL
  - Located
  - Near
- GENERAL AFFILIATION
  - Citizen-Resident-Ethnicity-Religion
- PART-WHOLE
  - Subsidiary
  - Geographical
- ORG AFFILIATION
  - Founder
  - Ownership
  - Membership
  - Sports-Affiliation
  - Investor
  - Student-Alum
  - Employment
- ARTIFACT
  - User-Owner-Inventor-Manufacturer

ACE corpus and competitions
ACE Relations Examples

Physical-Located PER-GPE
He was in Tennessee

Part-Whole-Subsidiary ORG-ORG
XYZ, the parent company of ABC

Person-Social-Family PER-PER
John’s wife Yoko

Org-AFF-Founder PER-ORG
Steve Jobs, co-founder of Apple...
Geographical Relations
Medical Relations

### UMLS Resource

<table>
<thead>
<tr>
<th>Injury</th>
<th>disrupts</th>
<th>Physiological Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bodily Location</td>
<td>location-of</td>
<td>Biologic Function</td>
</tr>
<tr>
<td>Anatomical Structure</td>
<td>part-of</td>
<td>Organism</td>
</tr>
<tr>
<td>Pharmacologic Substance</td>
<td>causes</td>
<td>Pathological Function</td>
</tr>
<tr>
<td>Pharmacologic Substance</td>
<td>treats</td>
<td>Pathologic Function</td>
</tr>
</tbody>
</table>
Medical Relations

Doppler echocardiography can be used to diagnose left anterior descending artery stenosis in patients with type 2 diabetes

↓

Echocardiography, Doppler DIAGNOSES Acquired stenosis
## Freebase Relations

Thousands of relations and millions of instances!
Manually created from multiple sources including Wikipedia InfoBoxes

<table>
<thead>
<tr>
<th>Relation name</th>
<th>Size</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>/people/person/nationality</td>
<td>281,107</td>
<td>John Dugard, South Africa</td>
</tr>
<tr>
<td>/location/location/contains</td>
<td>253,223</td>
<td>Belgium, Nijlen</td>
</tr>
<tr>
<td>/people/person/profession</td>
<td>208,888</td>
<td>Dusa McDuff, Mathematician</td>
</tr>
<tr>
<td>/people/person/place_of_birth</td>
<td>105,799</td>
<td>Edwin Hubble, Marshfield</td>
</tr>
<tr>
<td>/dining/restaurant/cuisine</td>
<td>86,213</td>
<td>MacAyo’s Mexican Kitchen, Mexican</td>
</tr>
<tr>
<td>/business/business_chain/location</td>
<td>66,529</td>
<td>Apple Inc., Apple Inc., South Park, NC</td>
</tr>
<tr>
<td>/biology/organism_classification_rank</td>
<td>42,806</td>
<td>Scorpaeniformes, Order</td>
</tr>
<tr>
<td>/film/film/genre</td>
<td>40,658</td>
<td>Where the Sidewalk Ends, Film noir</td>
</tr>
<tr>
<td>/film/film/language</td>
<td>31,103</td>
<td>Enter the Phoenix, Cantonese</td>
</tr>
<tr>
<td>/biology/organism_higher_classification</td>
<td>30,052</td>
<td>Calopteryx, Calopterygidae</td>
</tr>
<tr>
<td>/film/film/country</td>
<td>27,217</td>
<td>Turtle Diary, United States</td>
</tr>
<tr>
<td>/film/writer/film</td>
<td>23,856</td>
<td>Irving Shulman, Rebel Without a Cause</td>
</tr>
<tr>
<td>/film/director/film</td>
<td>23,539</td>
<td>Michael Mann, Collateral</td>
</tr>
<tr>
<td>/film/producer/film</td>
<td>22,079</td>
<td>Diane Eskenazi, Aladdin</td>
</tr>
<tr>
<td>/people/deceased_person/place_of_death</td>
<td>18,814</td>
<td>John W. Kern, Asheville</td>
</tr>
<tr>
<td>/music/artist/origin</td>
<td>18,619</td>
<td>The Octopus Project, Austin</td>
</tr>
<tr>
<td>/people/person/religion</td>
<td>17,582</td>
<td>Joseph Chartrand, Catholicism</td>
</tr>
<tr>
<td>/book/author/works_written</td>
<td>17,278</td>
<td>Paul Auster, Travels in the Scriptorium</td>
</tr>
<tr>
<td>/soccer/football_position/players</td>
<td>17,244</td>
<td>Midfielder, Chen Tao</td>
</tr>
<tr>
<td>/people/deceased_person/cause_of_death</td>
<td>16,709</td>
<td>Richard Daintree, Tuberculosis</td>
</tr>
<tr>
<td>/film/film/music</td>
<td>14,070</td>
<td>Stavisky, Stephen Sondheim</td>
</tr>
<tr>
<td>/business/company/industry</td>
<td>13,805</td>
<td>ATS Medical, Health care</td>
</tr>
</tbody>
</table>
Ontological Relations

IS-A (hypernym): subsumption between classes
- Giraffe IS-A ruminant IS-A ungulate IS-A mammal IS-A vertebrate IS-A animal...

Instance-of: relation between individual and class
- San Francisco instance-of city
"User defined" relations

Read Biology Papers → Extract Genes binding to Proteins

Read Web health Forums → Extract Drugs and reported side effects

Read CVEs → Extract vulnerable products, versions, exploit type

Read Nanorobotic papers → Extract nano-rod size, release mechanism, release time

Read Court Opinions → Extract Judge, case-type, holding, decision

Read News → Extract country1, country2, diplomatic event, time
Important property of RE

- **Long tail:**
  - Few "head relations" that many people care about.
  - Very many "tail relations" that are of interest to specific users.

- **Cannot assume a large training** set for each tail relation.

- **Domain knowledge** is important.
Relation Extraction vs. Question Answering

What are the similarities and differences between RE and QA?

[discuss]
Outline

Introduction to Relation Extraction

Hand-written Patterns

Supervised Machine Learning

Semi and Unsupervised Learning
Rules for IS-A Relation

Early intuition from Hearst (1992)

“Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use”

What does Gelidium mean?

How do you know?
Hearst’s Patterns for IS-A relations
Hearst’s Patterns for IS-A relations

“\(Y\) such as \(X ((, X)^* ((, \text{ and/or }) X))\)”
“such \(Y\) as \(X\)”
“\(X\) or other \(Y\)”
“\(X\) and other \(Y\)”
“\(Y\) including \(X\)”
“\(Y\), especially \(X\)”
Hearst’s Patterns for IS-A relations

“\( Y \) such as \( X ((, X)^* (, \text{ and} \mid \text{or}) X) \)”
“such \( Y \) as \( X \)”
“\( X \) or other \( Y \)”
“\( X \) and other \( Y \)”
“\( Y \) including \( X \)”
“\( Y \), especially \( X \)”
## Hearst’s Patterns for IS-A relations

<table>
<thead>
<tr>
<th>Hearst pattern</th>
<th>Example occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>X and other Y</td>
<td>temples, treasuries, <strong>and other</strong> important civic buildings.</td>
</tr>
<tr>
<td>X or other Y</td>
<td>Bruises, wounds, broken bones <strong>or other</strong> injuries...</td>
</tr>
<tr>
<td>Y such as X</td>
<td>The bow lute, <strong>such as</strong> the Bambara ndang...</td>
</tr>
<tr>
<td>Such Y as X</td>
<td><strong>such</strong> authors <strong>as</strong> Herrick, Goldsmith, and Shakespeare.</td>
</tr>
<tr>
<td>Y including X</td>
<td>common-law countries, <strong>including</strong> Canada and England...</td>
</tr>
<tr>
<td>Y, especially X</td>
<td>European countries, <strong>especially</strong> France, England, and Spain...</td>
</tr>
</tbody>
</table>
### Hearst’s Patterns for IS-A relations

<table>
<thead>
<tr>
<th>Hearst pattern</th>
<th>Example occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>X and other Y</td>
<td>...temples, treasuries, and other important civic buildings.</td>
</tr>
<tr>
<td>X or other Y</td>
<td>Bruises, wounds, broken bones or other injuries...</td>
</tr>
<tr>
<td>Y such as X</td>
<td>The bow lute, such as the Bambara ndang...</td>
</tr>
<tr>
<td>Such Y as X</td>
<td>...such authors as Herrick, Goldsmith, and Shakespeare.</td>
</tr>
<tr>
<td>Y including X</td>
<td>...common-law countries, including Canada and England...</td>
</tr>
<tr>
<td>Y, especially X</td>
<td>European countries, especially France, England, and Spain...</td>
</tr>
</tbody>
</table>
## Hearst’s Patterns for IS-A relations

<table>
<thead>
<tr>
<th>Hearst pattern</th>
<th>Example occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>X and other Y</td>
<td>...temples, treasuries, and other important civic buildings.</td>
</tr>
<tr>
<td>X or other Y</td>
<td>Bruises, wounds, broken bones or other injuries...</td>
</tr>
<tr>
<td>Y such as X</td>
<td>The bow lute, such as the Bambara ndang...</td>
</tr>
<tr>
<td>Such Y as X</td>
<td>...such authors as Herrick, Goldsmith, and Shakespeare.</td>
</tr>
<tr>
<td>Y including X</td>
<td>...common-law countries, including Canada and England...</td>
</tr>
<tr>
<td>Y, especially X</td>
<td>European countries, especially France, England, and Spain...</td>
</tr>
</tbody>
</table>
Hearst’s Patterns for IS-A relations

<table>
<thead>
<tr>
<th>Hearst pattern</th>
<th>Example occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>X and other Y</td>
<td>...temples, treasuries, and other important civic buildings.</td>
</tr>
<tr>
<td>X or other Y</td>
<td>Bruises, wounds, broken bones or other injuries...</td>
</tr>
<tr>
<td>Y such as X</td>
<td>The bow lute, such as the Bambara ndang...</td>
</tr>
<tr>
<td>Such Y as X</td>
<td>...such authors as Herrick, Goldsmith, and Shakespeare.</td>
</tr>
<tr>
<td>Y including X</td>
<td>...common-law countries, including Canada and England...</td>
</tr>
<tr>
<td>Y, especially X</td>
<td>European countries, especially France, England, and Spain...</td>
</tr>
</tbody>
</table>
## Hearst’s Patterns for IS-A relations

<table>
<thead>
<tr>
<th>Hearst pattern</th>
<th>Example occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>X and other Y</td>
<td>...temples, treasuries, <strong>and other</strong> important civic buildings.</td>
</tr>
<tr>
<td>X or other Y</td>
<td>Bruises, wounds, broken bones <strong>or other</strong> injuries...</td>
</tr>
<tr>
<td>Y such as X</td>
<td>The bow lute, <strong>such as</strong> the Bambara ndang...</td>
</tr>
<tr>
<td>Such Y as X</td>
<td><strong>...such</strong> authors <strong>as</strong> Herrick, Goldsmith, and Shakespeare.</td>
</tr>
<tr>
<td>Y including X</td>
<td>...common-law countries, <strong>including</strong> Canada and England...</td>
</tr>
<tr>
<td>Y, especially X</td>
<td>European countries, <strong>especially</strong> France, England, and Spain...</td>
</tr>
</tbody>
</table>
Patterns for learning meronyms

• Berland & Charniak (1999) tried it
• Selected initial patterns by finding all sentences in a corpus containing

  \textit{basement} and \textit{building}

\[
\begin{align*}
\text{whole } & \text{NN[-PL] 's POS part } \text{NN[-PL]} \quad \ldots \text{building's basement }...
\text{part } & \text{NN[-PL] of PREP \{the|a\} DET mods [JJ|NN]* whole } \text{NN} \quad \ldots \text{basement of a building }...
\text{part } & \text{NN in PREP \{the|a\} DET mods [JJ|NN]* whole } \text{NN} \quad \ldots \text{basement in a building }...
\text{parts } & \text{NN-PL of PREP wholes } \text{NN-PL} \quad \ldots \text{basements of buildings }...
\text{parts } & \text{NN-PL in PREP wholes } \text{NN-PL} \quad \ldots \text{basements in buildings }...
\end{align*}
\]

• Then, for each pattern:
  1. found occurrences of the pattern
  2. filtered those ending with \textit{-ing}, \textit{-ness}, \textit{-ity}
  3. applied a likelihood metric — poorly explained
• Only the first two patterns gave decent (though not great!) results
Intuition:

Relations often hold between specific types of entities

- located-in (ORGANIZATION, LOCATION)
- founded (PERSON, ORGANIZATION)
- cures (DRUG, DISEASE)

Start with Named Entity tags to extract relation!
Entity Types aren’t enough

Which relations hold between 2 entities?

Drug

Cure?

Prevent?

Cause?

Disease
Which relations hold between two entities?

PERSON

Founder?
Investor?
Member?
Employee?
President?

ORGANIZATION
Extracting Richer Relations Using Rules and Named Entities

Who holds what office in what organization?

PERSON,  POSITION of ORG
- George Marshall, Secretary of State of the United States

PERSON(named|appointed|chose|etc.)  PERSON Prep?  POSITION
- Truman appointed Marshall Secretary of State

PERSON [be]? (named|appointed|etc.) Prep?  ORG POSITION
- George Marshall was named US Secretary of State
Using Dependency Trees

Complex Surface Patterns

Combine tokens, dependency paths, and entity types to define rules.

Bill Gates, the CEO of Microsoft, said ...
Mr. Jobs, the brilliant and charming CEO of Apple Inc., said ...
... announced by Steve Jobs, the CEO of Apple.
... announced by Bill Gates, the director and CEO of Microsoft.
... mused Bill, a former CEO of Microsoft.
and many other possible instantiations...
Using Dependency Trees

Rule-Based Extraction

Use a collection of rules as the system itself

Source:
- Manually specified
- Learned from Data

Multiple Rules:
- Attach priorities/precedence
- Attach probabilities (more later)
Hearst’s Patterns for IS-A relations

“Y such as X ((, X)* (, and|or) X)”
“such Y as X”
“X or other Y”
“X and other Y”
“Y including X”
“Y, especially X”

Can also greatly benefit from using dependency trees!
(why?)

(is there something better than dependency trees?)
Hand-built patterns for relations

**Pluses**
- Human patterns tend to be high-precision
- Can be tailored to specific domains
- Easy to debug: why a prediction was made, how to fix?

**Minuses**
- Human patterns are often low-recall
- A lot of work to think of all possible patterns!
- Don’t want to have to do this for every relation!
- We’d like better accuracy (*generalization*)
Hand-built patterns for relations

Big research interests (for Yoav):

- How do we allow people to specify the relations they want?
- What is a good representation to work with?
- How can we improve the human pattern-writing process?
Outline

- Introduction to Relation Extraction
- Hand-written Patterns
- Supervised Machine Learning
- Semi and Unsupervised Learning
Supervised Machine Learning

Choose a set of relations we’d like to extract

Choose a set of relevant named entities

Find and label data
- Choose a representative corpus
- Label the named entities in the corpus
- Hand-label the relations between these entities
- Break into training, development, and test

Train a classifier on the training set
Automated Content Extraction

- **PERSON-SOCIAL**
  - Family
  - Business
  - Lasting Personal

- **PHYSICAL**
  - Near
  - Located

- **GENERAL AFFILIATION**
  - Citizen-Resident-Ethnicity-Religion

- **PART-WHOLE**
  - Subsidiary
  - Geographical

- **ORG AFFILIATION**
  - Founder
  - Ownership
  - Membership
  - Investor
  - Student-Alum
  - Employment
  - Sports-Affiliation

- **ARTIFACT**
  - User-Owner-Inventor-Manufacturer

ACE 2008 “Relation Extraction Task”

CS 295: STATISTICAL NLP (WINTER 2017)
Relation Extraction

Classify the relation between two entities in a sentence

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.
Word Features for Relation Extraction

*American Airlines*, a unit of AMR, immediately matched the move, spokesman *Tim Wagner* said

Mention 1

Mention 2

Headwords of M1 and M2, and combination
- Airlines
- Wagner
- Airlines-Wagner

Bag of words and bigrams in M1 and M2

\{American, Airlines, Tim, Wagner, American Airlines, Tim Wagner\}

Words or bigrams in particular positions left and right of M1/M2
- *M2: -1 spokesman*
- *M2: +1 said*

Bag of words or bigrams between the two entities

\{a, AMR, of, immediately, matched, move, spokesman, the, unit\}
American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.
Dependency Parse Features for Relation Extraction

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said

Base syntactic chunk sequence from one to the other

NP  NP  PP  VP  NP  NP

Constituent path through the tree from one to the other

NP  ↑  NP  ↑  S  ↑  S  ↓  NP

Dependency path

Airlines  matched  Wagner  said
Gazetteer and Trigger word features for relation extraction

Trigger list for family: kinship terms
- parent, wife, husband, grandparent, etc. [from WordNet]

Gazetteer:
- Lists of useful geo or geopolitical words
  - Country name list
  - Other sub-entities
American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

Entity-based features
- Entity\_1 type: ORG
- Entity\_1 head: airlines
- Entity\_2 type: PERS
- Entity\_2 head: Wagner
- Concatenated types: ORG\_PERS

Word-based features
- Between-entity bag of words: { a, unit, of, AMR, Inc., immediately, matched, the, move, spokesman }
- Word(s) before Entity\_1: NONE
- Word(s) after Entity\_2: said

Syntactic features
- Constituent path: \( NP \uparrow NP \uparrow S \uparrow S \downarrow NP \)
- Base syntactic chunk path: \( NP \rightarrow NP \rightarrow PP \rightarrow NP \rightarrow VP \rightarrow NP \rightarrow NP \)
- Typed-dependency path: \( Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner \)
Supervised Extraction

Machine Learning: hopefully, generalizes the labels in the right way

Use all of NLP as features: words, POS, NER, dependencies, embeddings

However

Usually, a lot of labeled data is needed, which is expensive & time consuming. Requires a lot of feature engineering!

P(birthplace) = 0.75

Classifer

Feature Engineering

John was born in [Liverpool], to Julia and Alfred Lennon.
Supervised Relation Extraction

**Pluses**
- Can get high accuracies if enough training data
- If test similar enough to training
- Can utilize a number of NLP tasks

**Minuses**
- Labeling a large training set is expensive
- Supervised models are brittle, don’t generalize well to different genres
Hand-built patterns for relations

Big research interests (for Yoav):

- Can we effectively learn from few examples?
- How do we effectively specify domain knowledge to an ML based system?
- Can we combine supervised ML and pattern-based RE? how?
- Can we improve a trained model by providing rules?
Outline

- Introduction to Relation Extraction
- Hand-written Patterns
- Supervised Machine Learning
- Semi and Unsupervised Learning
Seed-based or bootstrapping approaches to relation extraction

No training set? Maybe you have:

- A few seed tuples  or
- A few high-precision patterns

Can you use those seeds to do something useful?

- **Bootstrapping**: use the seeds to directly learn a relation
Relation Bootstrapping

Gather a set of seed pairs that have the relation
1. Find sentences with these pairs
2. Look at the context between or around the pair and generalize the context to create patterns
3. Use the patterns to gather more pairs
4. Repeat
Bootstrapping Example

<Mark Twain, Elmira> Seed tuple od “died in”

Look for the environments of the seed tuple

“Mark Twain is buried in Elmira, NY.”
X is buried in Y

“The grave of Mark Twain is in Elmira”
The grave of X is in Y

“Elmira is Mark Twain’s final resting place”
Y is X’s final resting place.

Use those patterns to find new tuples

Repeat
Dipre: Extract <author,book> pairs

Start with 5 seeds:

<table>
<thead>
<tr>
<th>Author</th>
<th>Book</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isaac Asimov</td>
<td>The Robots of Dawn</td>
</tr>
<tr>
<td>David Brin</td>
<td>Startide Rising</td>
</tr>
<tr>
<td>James Gleick</td>
<td>Chaos: Making a New Science</td>
</tr>
<tr>
<td>Charles Dickens</td>
<td>Great Expectations</td>
</tr>
<tr>
<td>William Shakespeare</td>
<td>The Comedy of Errors</td>
</tr>
</tbody>
</table>

Find Instances on the Web:

- The Comedy of Errors, by William Shakespeare, was
- The Comedy of Errors, by William Shakespeare, is
- The Comedy of Errors, one of William Shakespeare's earliest attempts
- The Comedy of Errors, one of William Shakespeare's most

Extract patterns (group by middle, take longest common prefix/suffix)

?x, by ?y, ?x, one of ?y’s

Now iterate, finding new seeds that match the pattern
Snowball

Similar iterative algorithm

Group instances w/similar prefix, middle, suffix, extract patterns
  - But require that X and Y be named entities
  - And compute a confidence for each pattern

<table>
<thead>
<tr>
<th>Organization</th>
<th>Location of Headquarters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft</td>
<td>Redmond</td>
</tr>
<tr>
<td>Exxon</td>
<td>Irving</td>
</tr>
<tr>
<td>IBM</td>
<td>Armonk</td>
</tr>
</tbody>
</table>

.69  ORGANIZATION  {‘s, in, headquarters}  LOCATION

.75  LOCATION  {in, based}  ORGANIZATION
Distant Supervision

Combine bootstrapping with supervised learning

- Instead of 5 (or just a few) seeds,
  - Use a large database to get huge # of seed examples
- Create lots of features from all these examples
- Combine in a supervised classifier
Distantly Supervised learning of relation extraction patterns

1. For each relation

2. For each tuple in big database

3. Find sentences in large corpus with both entities

4. Extract frequent features (parse, words, etc)

5. Train supervised classifier using these patterns

\[
P(\text{born-in} \mid f_1, f_2, f_3, \ldots, f_{70000})
\]

Born-In

<Edwin Hubble, Marshfield>
<Albert Einstein, Ulm>

Hubble was born in Marshfield
Einstein, born (1879), Ulm
Hubble’s birthplace in Marshfield

PER was born in LOC
PER, born (XXXX), LOC
PER’s birthplace in LOC
Distant Supervision Paradigm

Like supervised classification:
- Uses a classifier with lots of features
- Supervised by detailed hand-created knowledge
- Doesn’t require iteratively expanding patterns

Like unsupervised classification:
- Uses very large amounts of unlabeled data
- Not sensitive to genre issues in training corpus
Unsupervised Relation Extraction

Open Information Extraction:
- extract relations from the web with no training data, no list of relations

1. Use parsed data to train a “trustworthy tuple” classifier
2. Single-pass extract all relations between NPs, keep if trustworthy
3. Assessor ranks relations based on text redundancy

- (FCI, specializes in, software development)
- (Tesla, invented, coil transformer)
- (Tesla, inventor of, transformer)
Evaluation of Semi-supervised and Unsupervised Relation Extraction

Since it extracts totally new relations from the web
  ◦ There is no gold set of correct instances of relations!
  ◦ Can’t compute precision (don’t know which ones are correct)
  ◦ Can’t compute recall (don’t know which ones were missed)

Instead, we can approximate precision (only)
  ◦ Draw a random sample of relations from output, check precision manually

\[
\hat{P} = \frac{\text{# of correctly extracted relations in the sample}}{\text{Total # of extracted relations in the sample}}
\]

Can also compute precision at different levels of recall.
  ◦ Precision for top 1000 new relations, top 10,000 new relations, top 100,000
  ◦ In each case taking a random sample of that set

But no way to evaluate recall
Outline

Introduction to Relation Extraction

Hand-written Patterns

Supervised Machine Learning

Semi and Unsupervised Learning
Relation and Event Extraction

Perhaps the most important application of practical NLP

Can do many useful things

Far from being a solved problem

Touches almost all areas of language processing